

We claim:

1 1. A server-side caching method for use with a server cluster
2 including at least one central storage device and a plurality of servers having
3 respective cache storage devices, the method comprising the steps of:
4 receiving a client request for a data object from a client device with
5 one of the servers;
6 determining whether the data object is being cached by the server
7 that received the client request;
8 determining whether a server that did not receive the client
9 request is caching the data object in response to a determination that the server
10 that received the client request is not caching the data object;
11 obtaining a copy of the data object from the cache storage device
12 of a server that did not receive the client request in response to a determination
13 that the server that received the client request is not caching the data object and
14 a determination that the data object is cached in a server that did not receive the
15 client request; and
16 transmitting the data object to the client device.

1 2. A method as claimed in claim 1, wherein the step of transmitting
2 the data object to the client device comprises:
3 transmitting the data object from the server that received the client
4 request to the client device in response to a determination that the data object is
5 being cached by the server that received the client request.

1 3. A method as claimed in claim 1, wherein the step of transmitting
2 the data object to the client device comprises:
3 sending the copy of the data object from the cache storage
4 device of a server that did not receive the client request to the server that
5 received the client request; and
6 sending the copy of the data object from the server that received
7 the client request to the client device.

1 4. A method as claimed in claim 3, further comprising the step of:
2 sending at least one of data object request rate information and
3 a data object replication recommendation to the server that received the client
4 request with the copy of the data object.

1 5. A method as claimed in claim 1, wherein the step of transmitting
2 the data object to the client device comprises:
3 sending a copy of the data object from the central storage device
4 to the server that received the client request in response to a determination that
5 the server that received the client request is not caching the data object and a
6 determination that the data object is not cached in a server that did not
7 receive the client request.

1 6. A method as claimed in claim 5, further comprising the step of:
2 caching the copy of the data object in the cache storage device of
3 the server that received the client request.

1 7. A method as claimed in claim 1, further comprising the step of:
2 maintaining a cache status table including a list of data objects
3 stored in the respective cache storage devices of the plurality of servers.

1 8. A method as claimed in claim 7, wherein the step of determining
2 whether a server that did not receive the client request is caching the data
3 object comprises:
4 querying the cache status table.

1 9. A method as claimed in claim 7, wherein the step of transmitting
2 the data object to the client device comprises:
3 sending a copy of the data object from the central storage device
4 to the server that received the client request in response to a determination that
5 the server that received the client request is not caching the data object and a
6 determination that the data object is not listed in cache status table.

1 10. A method as claimed in claim 9, further comprising the steps of:

2 caching the data object in the cache storage device of the server
3 that received the client request; and
4 updating the cache status table to reflect that a copy of the data
5 object has been cached in the server that received the client request.

1 11. A method as claimed in claim 1, wherein the step of obtaining a
2 copy of the data object from the cache storage device of a server that did not
3 receive the client request comprises:

4 obtaining a copy of the data object from the cache storage
5 device of a server is operating below a predetermined load threshold.

1 12. A server system for use with a plurality of client devices,
2 comprising;

3 a local area network adapted to be connected to a wide area
4 network;

5 a plurality of servers connected to the local area network and
6 having respective cache storage devices adapted to cache data objects; and

7 a cache load server process running on at least one of the
8 servers that maintains a list of data objects cached in the cache storage
9 devices.

1 13. A server system as claimed in claim 12, wherein each of the
2 servers runs a load daemon process that monitors server loading and
3 transmits server loading information to the cache load server process.

1 14. A server system as claimed in claim 12, wherein each of the
2 servers runs a web server process that receives requests for data objects
3 from client devices and queries the cache load server process to determine
4 whether another server is caching a requested data object in response to a
5 determination that the associated server is not caching the requested data
6 object.

1 15. A server system as claimed in claim 14, wherein the web server
2 processes queries the cache load server process to determine whether a

3 server that is caching a requested data object is below a predetermined
4 server loading threshold.

1 16. A server system as claimed in claim 12, further comprising:
2 a dispatcher adapted to connect the local area network to the
3 wide area network.

1 17. A server system as claimed in claim 12, further comprising:
2 at least one central storage device connected to the local area
3 network.

1 18. A server system for use with a plurality of client devices,
2 comprising;
3 a local area network adapted to be connected to a wide area
4 network;
5 a plurality of servers connected to the local area network and
6 having respective cache storage devices adapted to cache data objects; and
7 means for cooperatively caching data objects within the cache
8 storage devices such that cached data objects may be shared by the plurality of
9 servers.

1 19. A server system as claimed in claim 18, wherein each of the
2 servers includes means for receiving a client request for a data object from a
3 client device and means for determining whether the data object is being
4 cached locally.

1 20. A server system as claimed in claim 18, wherein the means for
2 cooperatively caching data objects includes means for determining whether a
3 server that did not receive the client request is caching the data object.

1 21. A server system as claimed in claim 20, wherein the means for
2 cooperatively caching data objects includes means for determining whether a
3 server that did not receive the client request and is caching the data object is
4 operating below a predetermined loading threshold.

[illegible]